XplainableClusterExplorer: A novel approach for Interactive Feature Selection for Clustering

Eric Fezer
Aalen University of Applied Sciences
73430 Aalen, Germany

Dominik Raab
Aalen University of Applied Sciences
73430 Aalen, Germany

Andreas Theissler
Aalen University of Applied Sciences
73430 Aalen, Germany
orcid.org/0000-0003-0746-0424

ABSTRACT

Human-centered machine learning is becoming an emerging field aiming to enable domain experts that do not necessarily have a data science background to make use of machine learning applications. Especially in unsupervised machine learning, e.g. cluster analysis, models cannot be autonomously tuned towards an optimal solution for a given application due to the absence of ground truth like class labels. In cluster analysis, different feature subsets may lead to different clusterings. The identification of the best subset of given features is therefore essential in order to improve the overall clustering performance and to obtain a clustering that is suitable for a given application. To support users in finding an optimal clustering solution, we propose XplainableClusterExplorer, an interactive and explorative approach suitable for feature selection for clustering. In an interactive combination of user and machine learning models, the user is supported by evaluation criteria and visualizations in determining feature subsets and adjusting hyperparameters. For feature subset selection we propose a combination with feature importances from random forests and LIME. Since this requires a supervised setting, the cluster assignments are used as tentative class labels in subsequent step. Our experimental results have shown that this subsequent classification step leveraging calculated feature importances can facilitate feature subset selection and therefore enhance overall clustering performance.

CCS CONCEPTS

• **Computing methodologies** → *Feature selection*; • **Human-centered computing** → **Visual analytics**.

KEYWORDS

Human-Centered Machine Learning, Unsupervised Machine Learning, Clustering, k-Means, Feature Selection

ACM Reference Format:
Eric Fezer, Dominik Raab, and Andreas Theissler. 2020. XplainableClusterExplorer: A novel approach for Interactive Feature Selection for Clustering. In *The 13th International Symposium on Visual Information Communication and Interaction (VINCI 2020), December 8–10, 2020, Eindhoven, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3430036.3430066

VINCI 2020, December 8–10, 2020, Eindhoven, Netherlands

ACM ISBN 978-1-4503-8750-7/20/12…$15.00
https://doi.org/10.1145/3430036.3430066

1 INTRODUCTION

The task of finding meaningful clusters is regularly encountered in data science projects, since in applications supervised machine learning (ML) can often not be applied due to the absence of labels. In fact, clustering can be a starting point to move towards supervised methods like classification. As clustering methods traditionally assume all features to be equally important, selecting a subset of important features is crucial [4]. The challenge is to determine features that maximally enhance the clustering performance [12] and to remove irrelevant features. Keeping non-informative features within the data set can lead to weak results, consumes more storage facilities and computing time [1].

To address this, we introduce an interactive and explorative approach for feature selection for clustering. In an interplay of user with both an unsupervised and a supervised ML model, feature subsets and model hyperparameters are determined in order to yield a clustering that is meaningful for a given problem setting. The main contributions of this short paper are:

(1) An interactive and explorative approach for obtaining a clustering suitable for a given application supported by evaluation criteria and visualizations.
(2) For feature subset selection a subsequent classification step is proposed, with identified clusters used as tentative classes in order to determine feature importances using LIME and random forests.
(3) Executable prototype: fezerraab.shinyapps.io/XplainClustExpl/

2 RELATED WORK

The general benefits of combining ML models with user interactions was for example shown by Holzinger et al. in [13].

For supervised ML, Theissler et al. [20] proposed an interactive approach to compare classifiers and Grimmeisen et al. an approach to interactively label data [10]. For anomaly detection, an interactive labelling process was proposed by Theissler et al. in [19].

Aiming to interactively and automatically identify interesting multi-dimensional subsets, Guo [11] developed a human-centered approach for interactive feature selection and multivariate hierarchical clustering based on computational and visual techniques. Dy and Brodley [8] designed a visual feature subset selection approach using expectation-maximization clustering algorithm.

To address the challenge of large numbers of feature subsets in high dimensional data, Goil et al. [9] presented MAFIA, an approach with adaptive grids for fast subspace clustering. Yuan et al. [23] suggest an approach based on visual exploration to either algorithmically or manually inspect data dimensions. Wang et al. [22] decompose high dimensional data into a continuum of generalized

3D subspaces. In order to make cluster assignments more traceable Dasgupta et al. [3] have developed an algorithm that generates explainable clusters using a tree with k leaves.

In contrast to the discussed work, we (1) guide the user in determining hyperparameters and finding an optimal feature subset, (2) enable the user to make use of his/her domain knowledge in order to investigate the clustering results with interactive visualizations, and (3) implement feature selection using feature importances by using identified clusters as tentative class labels in a subsequent classification step. The entire process is an interplay of user and ML model whereby the user is in-the-loop.

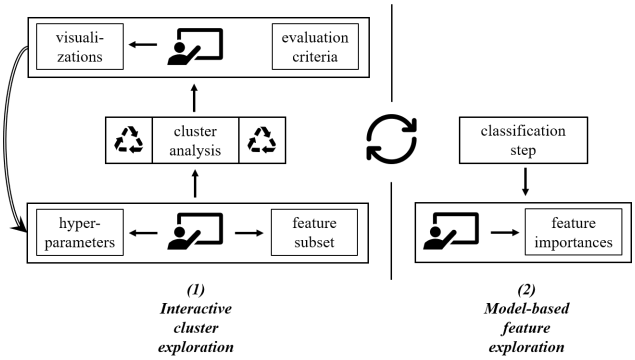

Figure 1: Our approach: with interactive visualizations and evaluation criteria in step (1) the user finds the desired clustering by determining hyperparameters and a feature subset. Using feature importances in step (2) the user utilizes these insights and adapts step (1) with adjusted settings.

3 THE APPROACH: INTERACTIVE FEATURE SELECTION FOR CLUSTERING

Our approach provides decision support in selecting a feature subset (see Definition (1)) in order to find a clustering (see Definition (2)) that is best for a given problem setting. In the prototype, the k-Means clustering algorithm is used. The ideas can, however, be transferred to alternative clustering methods.

Definition 1 (feature subset). *In a feature space $F = \{f_1, ..., f_N\}$ with N features (variables, attributes), a* feature subset F_i *is a subset of selected features, i.e. $F_i \subseteq F$.*

Definition 2 (clustering). *In a data set D with feature space F, a* clustering C_i *is the assignment of data points to k groups with $C_i = \{c_1, ..., c_k\}$. C_i is determined by the clustering method's hyperparameters Φ and the selected feature subset F_i, i.e. $C_i = f(F_i, \Phi)$.*

The goal was to develop an approach with a novel step for interactive feature selection for cluster analysis applicable by domain experts and data scientists. Since domain experts do not necessarily have a data science background, a broad variety of self-descriptive and easily interpretable visualizations in form of 2D and 3D scatter plots, parallel coordinates plot and correlograms are used. Selecting or deselecting features leads to an immediate adjustment of the visualizations and evaluation parameters. XplainableClusterExplorer is implemented in R [15] with shiny and plotly [14] and is accessible online using own data sets. The approach is depicted in Fig. 1. It comprises two essential steps where in both the user is in-the-loop: (1) interactive cluster exploration and (2) model-based feature exploration.

3.1 Step 1: Interactive cluster exploration

In step 1 (*interactive cluster exploration*), the clustering methods' hyperparameters Φ are suggested and chosen by the user and the optimal feature subset F_i is selected, both supported by evaluation metrics (see eq. (1) and eq. (2)). During this step, the user interacts with several interactive visualizations and either relies on his/her domain knowledge or executes a structured or random search to find the optimal feature subset leading to the best clustering performance. Promising candidates of (F_i, Φ) can be stored. In addition, the set of (F_i, Φ) that achieved the best clustering is automatically stored. This allows for later comparison as well as for the reproduction of clustering results.

3.2 Step 2: Model-based feature exploration

In a feature space F with N features, the number of possible feature subsets is $2^N - 1$ (excluding the empty set). Even for moderate N it becomes infeasible to manually explore all feature subsets. While domain knowledge is expected to guide the user and hence reduce the search space, the support of supervised models is proposed yielding a ranking of the current feature subset F_i.

To achieve this, in step 2 (*model-based feature exploration*), we enhance our approach by a subsequent classification step where the identified clusters are used as tentative class labels. This enables the user to make use of feature importances calculated by both random forests and LIME. With these insights the user can adjust the feature selection in step 1 and revise the feature subset. As the settings in step 1 influence the insights generated in step 2 and these in turn affect the feature subset selection in step 1, this forms a loop of user and ML models allowing for iterative improvement of results leveraging the strengths of models and users.

Based on established cluster evaluation criteria, the user can interactively detect the feature subsets F_i that lead to the best clustering. In addition, the number of clusters is evaluated with metrics. Furthermore, XplainableClusterExplorer uses a broad variety of interactive visualizations that enable users to (1) inspect correlations between the features, (2) determine the number of clusters, (3) examine high-dimensional data using parallel coordinates plot, (4) investigate the clustering results with a 2D and 3D scatter plot, (5) compare the current settings with the best cluster result achieved so far, and (6) store promising clusterings.

In contrast to using traditional clustering evaluation criteria, we transfer the successfully applied feature selection in supervised ML based on feature importances to clustering methods.

4 XPLAINABLECLUSTEREXPLORER

4.1 Interactive cluster exploration pane

On a side panel on the left, the user both can set the clustering method's hyperparameters (in our prototype the number of clusters for the k-Means algorithm) and select the feature subset F_i. Changes instantly trigger recalculation of the clustering C_i and as a result of the interactive visualizations.

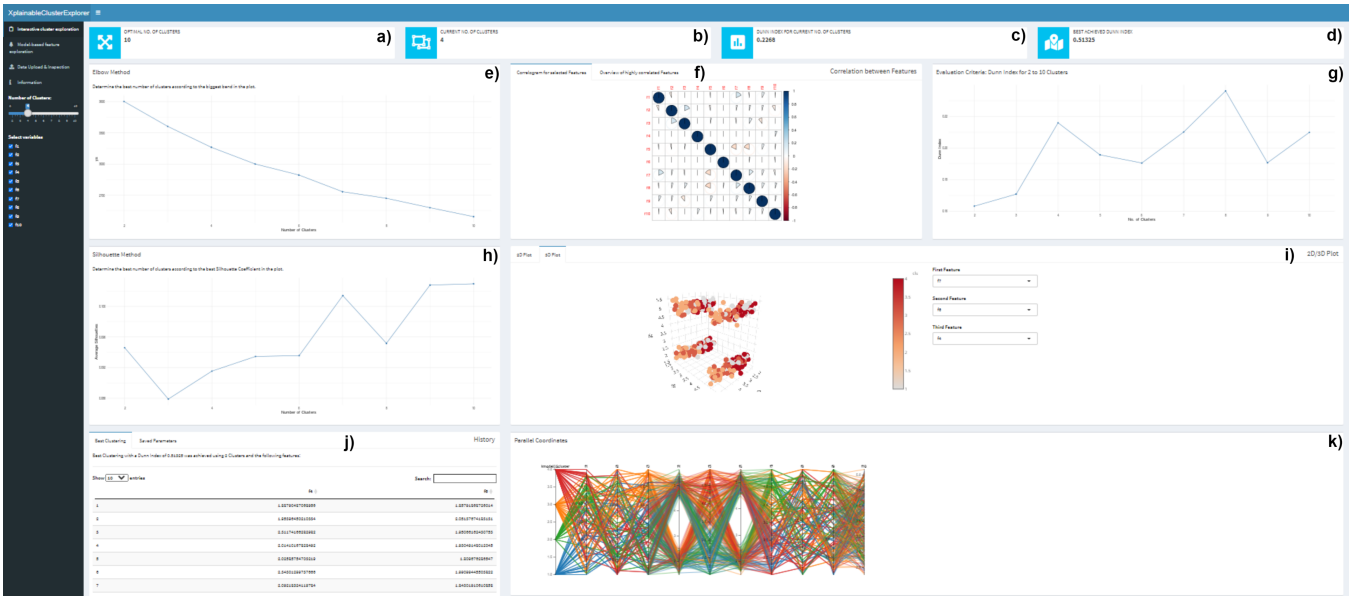

Figure 2: Interactive cluster exploration pane, to determine best no. of clusters and optimal feature subset based on evaluation criteria and interactive visualizations.

The *Interactive cluster exploration pane* (Fig. 2) is comprised of the following components:

- *Optimal No. of Clusters:* This infobox (Fig. 2, a) provides a recommendation for setting the optimal number of clusters based on the silhouette coefficient. The silhouette coefficient [17] functions as a criterion for estimating the optimal number of clusters and is calculated for each data instance x_m using the mean intra-cluster distance $dist(x_m - c_i)$ and mean nearest-cluster distance $dist(x_m - c_j)$, where c_i is the cluster x_m belongs to and c_j is the neighboring cluster. The silhouette coefficient expresses how strongly x_m belongs to its assigned cluster c_i and to what extent it could belong to its neighboring cluster c_j. Afterwards, the mean silhouette coefficient is computed to evaluate the optimal number of clusters. The equation of the silhouette coefficient for a single data instance x_m is

$$S(x_m) = \frac{dist(x_m - c_i) - dist(x_m - c_j)}{max(dist(x_m - c_i), dist(x_m - c_j))} \quad (1)$$

  which is then summed up for all instances of all clusters.
- *Current No. of Clusters:* This infobox (Fig. 2, b) shows the number of clusters the user has selected.
- *Dunn Index for Current No. of Clusters | Best Achieved Dunn Index:* The two infoboxes (Fig. 2, c & d) on the top right corner show the Dunn Index, a metric to evaluate the quality of clustering [7]. The metric itself is an internal evaluation scheme aiming to identify clusters having both a small variance between the data points of their associated clusters and having the means of the different clusters far apart compared to the variance within the k clusters. It is defined as:

$$DunnIndex = \min_{(1 \le i \le k)} \{ \min_{1 \le j \le k, j \ne i} \{ \frac{\delta(c_i, c_j)}{\max\{(\Delta(c_m))\}} \} \} \quad (2)$$

  where, $\delta(c_i, c_j)$ is the inter-cluster distance, e.g. the distance between the clusters c_i and c_j, and $\Delta(c_m)$ the intra-cluster distance, which is the distance within the cluster c_m. The infobox (c) shows the Dunn Index for the current combination of features and the number of clusters, while the infobox (d) displays the current best achieved Dunn Index.
- *Elbow Method:* Besides the silhouette coefficient, the elbow method is a popular approach specifying the optimal number of clusters within a data set. The elbow method is based on a visual inspection of finding an 'elbow' in the plot (Fig. 2, e) and picking the underlying number of clusters as the best number of clusters within the data set. Therefore, the within-cluster sum of squared errors is calculated for two to ten clusters and shown in a line chart.
- *Correlogram:* The features of the selected feature subset F_i are shown in a correlation matrix (Fig. 2, f). Pearson's R is visualized in a pie chart for each possible feature correlation. This allows to identify linearly correlated features as well as linearly unrelated features – both could be candidates to be removed. In addition, an overview of highly correlating features (|Pearson's R| > 0.7) is available.
- *Evaluation Criteria: Dunn Index for 2 to 10 Clusters:* This line plot (Fig. 2, g) shows a comparison of the Dunn indices of the current feature subset for k-Means with two to ten clusters.
- *Silhouette Method:* This line plot (Fig. 2, h) contains the calculated average silhouette coefficients for two to ten clusters enabling the user to compare the calculated clustering performance for various numbers of clusters. The higher the

average silhouette coefficient the better the number of clusters for the k-Means algorithm.
- *Interactive 2D-Plot | Features for 2D-Plot:* Here, the user can select features to plot in a 2D scatter plot (Fig. 2, i). Data points are colored according to their cluster assignment and the respective cluster centers are depicted by an 'X'.
- *Interactive 3D-Plot | Features for 3D-Plot:* In this visualization (Fig. 2, i) the user can compare three features using a 3D scatter plot[1] enabling the user to (1) dynamically rotate the perspective on the data points, (2) display the respective feature values by clicking on the single data points, and (3) obtain an overview over the spatial separation of the data points and then compare that spatial separation with a different combination of features selected. In addition, data points are colored according to their cluster assignment.
- *History:* In the history (Fig. 2, j) the best clustering are stored, determined by the Dunn Index. Additionally, the user can save cluster settings. The number of clusters $|C_i|$ and the feature subset F_i are stored. This allows to reproduce a particular clustering and compare different settings.
- *Parallel Coordinates:* In this plot (Fig. 2, k) an axis for each selected feature is drawn. We decided to use a parallel coordinates plot, since it allows for a quick overview over the attributes' distribution and allows to rapidly identify clusters within a specific feature.

4.2 Model-based feature exploration pane

The *Model-based feature exploration pane* contains the following components:

- *Accuracy of Random Forests | Sensitivity of Random Forests | Specificity of Random Forests:* These infoboxes show the respective metrics determined by a random forest [2] with 10-fold cross-validation and a train-test-ratio of 75/25.
- *Class Distribution using Clusters as Classes:* This plot shows the number of data points per cluster based on the assignment of k-Means in order to evaluate cluster sizes.
- *Random Forests' Confusion Matrix:* Here, the confusion matrix of the RF is displayed based on a 10-fold cross-validation and a train-test-ratio of 75/25. The identified clusters are used as tentative classes.
- *LIME Feature Importance:* By applying the model interpretability package Local Interpretable Model-agnostic Explanations (LIME) that is used to explain the output of any ML model based on local surrogate models [16], the user can observe the RF-based feature importances for an accurate classification. Therefore, randomly selected data points are shown for feature importance investigation.
- *Random Forests' Variable Importance:* Using random forests variable importance[2], the most important features can be extracted and visualized in form of a bar plot. The higher the variable importance, the more informative the feature and the stronger its impact on the model's output.

[1]The 3D scatter plot from the R package 'plotly' was used.
[2]random forest of the 'caret' package was used.

5 WALKTHROUGH AND EVALUATION

With the following walkthrough[3], we seek to illustrate the utility of XplainableClusterExplorer in order to find the hyperparameters and feature subset that lead to the best clustering. As a preliminary validation of the approach, an artificial data set with an unknown number of hidden in a subspace of a 10-dimensional feature space ($f_1 ... f_{10}$) was created by this paper's third author. The first and second author used XplainableClusterExplorer to detect these clusters.

From a user's point of view, determining the optimal number of clusters is an essential step, as this highly influences the clustering outcome. In finding the optimal number of clusters, XplainableClusterExplorer supports us by means of the elbow method and the silhouette method plot. For the initial feature selection containing all features, according to the elbow method plot's biggest bend, the optimal number of clusters seems to be six. After initially adjusting the number of clusters, we seek to deselect non-informative and redundant features. A powerful plot for deselecting features is the parallel coordinates plot. Here, features *f4*, *f5* and *f6* separate the data instances much better than the remaining features, which we subsequently deselect for this reason. Afterwards, we switch to the *model-based feature exploration* pane in order to investigate the feature importances. According to feature importances of both the random forests and LIME, *f4* and *f6* tends to be more important than *f5*, hence, we deselected *f5*. This alters the suggested number of clusters. Both the elbow method and the silhouette method plot indicate four clusters. Additionally considering the Dunn Index, we set the number of clusters to four. This leads to the best Dunn Index achieved so far. Finally, by visually investigating the current clustering with the 2D-/3D-scatter plot and the parallel coordinates plot we can ensure that the results are valid. The four clusters were hidden in the subspace of *f4* and *f6*.

In order to ensure our approach fulfills requirements for processing real world data sets, we evaluated XplainableClusterExplorer with two well-known data sets suitable for clustering identification, namely the Anuran Calls (MFCCs) data set [6] and the Cervical Cancer Behavior Risk Data Set [18]. Though, clusters were successfully isolated, performance issues occurred on the free hosting platform. This can, however, be solved running the R code locally.

6 CONCLUSION AND FUTURE WORK

The selection of informative features is necessary to obtain meaningful clusters. We proposed an interactive and explorative approach for feature selection. In a *interactive cluster exploration* step, the user is supported by visualizations and evaluation criteria followed by a *model-based feature exploration* step, where LIME and random forests are used to suggest informative features.

Future work could be to use of additional clustering algorithms with approaches such as hierarchical or fuzzy clustering. Furthermore, alternative metrics for evaluating the quality of clusters, such as the Davies-Bouldin Index [5] and GAP Statistic [21], could be added. We also plan to conduct user studies in order to evaluate the various functions of interactive feature selection. Additionally, we want to implement a function conducting an automatic feature subset creation and evaluation. Furthermore, we plan to add further visualizations and revise current visualizations.

[3]Video of walkthrough: https://youtu.be/5waQDul_L_4

REFERENCES

[1] Salem Alelyani, Jiliang Tang, and Huan Liu. 2013. Feature Selection for Clustering: A Review. In *Data Clustering*, Charu C. Aggarwal and Chandan K. Reddy (Eds.). Chapman and Hall/CRC, Boca Raton, FL, 29–60. https://doi.org/10.1201/9781315373515-2

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Sanjoy Dasgupta, Nave Frost, Michal Moshkovitz, and Cyrus Rashtchian. 2020. Explainable k-Means and k-Medians Clustering. *arXiv preprint arXiv:2002.12538* (2020).

[4] Manoranjan Dash, Kiseok Choi, Peter Scheuermann, and Huan Liu. 2002. Feature Selection for Clustering – A Filter Solution. (2002).

[5] David L. Davies and Donald W. Bouldin. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1, 2 (1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[6] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[7] J. C. Dunn. 1973. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* 3, 3 (1973), 32–57. https://doi.org/10.1080/01969727308546046

[8] Jennifer G. Dy and Carla E. Brodley. 2000. Visualization and interactive feature selection for unsupervised data. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, Raghu Ramakrishnan (Ed.). ACM, New York, NY, 360–364. https://doi.org/10.1145/347090.347168

[9] Sanjay Goil, Harsha Nagesh, and Alok Choudhary. 1999. MAFIA: Efficient and scalable subspace clustering for very large data sets. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 443. ACM, 452.

[10] Benedikt Grimmeisen and Andreas Theissler. 2020. The Machine Learning Model as a Guide: Pointing Users to Interesting Instances for Labeling through Visual Cues.. In *The 13th International Symposium on Visual Information Communication and Interaction (VINCI 2020), December 8–10,2020, Eindhoven, Netherlands*. ACM.

[11] Diansheng Guo. 2003. Coordinating Computational and Visual Approaches for Interactive Feature Selection and Multivariate Clustering. *Information Visualization* 2, 4 (2003), 232–246. https://doi.org/10.1057/palgrave.ivs.9500053

[12] Emrah Hancer, Bing Xue, and Mengjie Zhang. 2020. A survey on feature selection approaches for clustering. *Artificial Intelligence Review* 53, 6 (2020), 4519–4545. https://doi.org/10.1007/s10462-019-09800-w

[13] Andreas Holzinger. 2015. Interactive Machine Learning (iML). *Informatik-Spektrum* 39 (2015). https://doi.org/10.1007/s00287-015-0941-6

[14] Plotly Technologies Inc. 2015. *Collaborative data science*. Montreal, QC. https://plot.ly

[15] R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ ISBN 3-900051-07-0.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. [n.d.]. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://arxiv.org/pdf/1602.04938

[17] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[18] Sobar, Rizanda Machmud, and Adi Wijaya. 2016. Behavior Determinant Based Cervical Cancer Early Detection with Machine Learning Algorithm. *Advanced Science Letters* 22, 10 (2016), 3120–3123. https://doi.org/10.1166/asl.2016.7980

[19] Andreas Theissler, Anna-Lena Kraft, Max Rudeck, and Fabian Erlenbusch. 2020. VIAL-AD: Visual Interactive Labelling for Anomaly Detection – An approach and open research questions. In *4th International Workshop on Interactive Adaptive Learning (IAL2020)*. CEUR-WS.

[20] Andreas Theissler, Simon Vollert, Patrick Benz, Laurentius A Meerhoff, and Marc Fernandes. 2020. ML-ModelExplorer: An Explorative Model-Agnostic Approach to Evaluate and Compare Multi-class Classifiers. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 281–300.

[21] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 2 (2001), 411–423. https://doi.org/10.1111/1467-9868.00293

[22] B. Wang and K. Mueller. 2018. The Subspace Voyager: Exploring High-Dimensional Data along a Continuum of Salient 3D Subspaces. *IEEE Transactions on Visualization and Computer Graphics* 24, 2 (2018), 1204–1222.

[23] X. Yuan, D. Ren, Z. Wang, and C. Guo. 2013. Dimension Projection Matrix/Tree: Interactive Subspace Visual Exploration and Analysis of High Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2625–2633.